(12) UK Patent Application (19) GB (11) 2 347 053 (13) A

(43) Date of A Publication 23.08.2000

(21) Application No 9903672.5

(22) Date of Filing 17.02.1999

(51) INT CL[7]
H04L 12/58 , G06F 17/60

(52) UK CL (Edition R)
H4P PPA PX
U1S S2124

(56) Documents Cited
GB 2328110 A EP 0886228 A2 EP 0838774 A2
EP 0813162 A2 WO 98/37680 A2

(58) Field of Search
UK CL (Edition Q) H4P PPA PX
INT CL[6] G06F 17/60 , H04L 12/58
Online:WPI,EPODOC,JAPIO

(71) Applicant(s)
Argo Interactive Limited
(Incorporated in the United Kingdom)
7 Dukes Court, CHICHESTER, West Sussex,
PO19 2FX, United Kingdom

(72) Inventor(s)
Richard Jelbert
Jason Paul Tribbeck

(74) Agent and/or Address for Service
D Young & Co
21 New Fetter Lane, LONDON, EC4A 1DA,
United Kingdom

(54) Abstract Title
Proxy server filters unwanted email

(57) Email messages transmitted from a server via a mail transport protocol over an email network are passed through a proxy host 301, which is able to locally filter useful email from junk email by utilising a series of "scoring" metrics or by more explicit user configuration (killfiles), before passing the filtered mail on to the client user via the chosen mail transport protocol. The proxy server can produce logs and digests of processed junk email and send them by email or present them via a secure World Wide Web document to a system administrator for inspection, and also, for any message which cannot be conclusively scored as being junk, add it to a second per-user mailbox which can be inspected by the intended recipient at his discretion via a World Wide Web interface. The user can then inform the proxy definitively by World Wide Web fill-out form whether the message is junk or not, and email messages confirmed as being junk can then be removed automatically from all mailboxes held on the proxy.

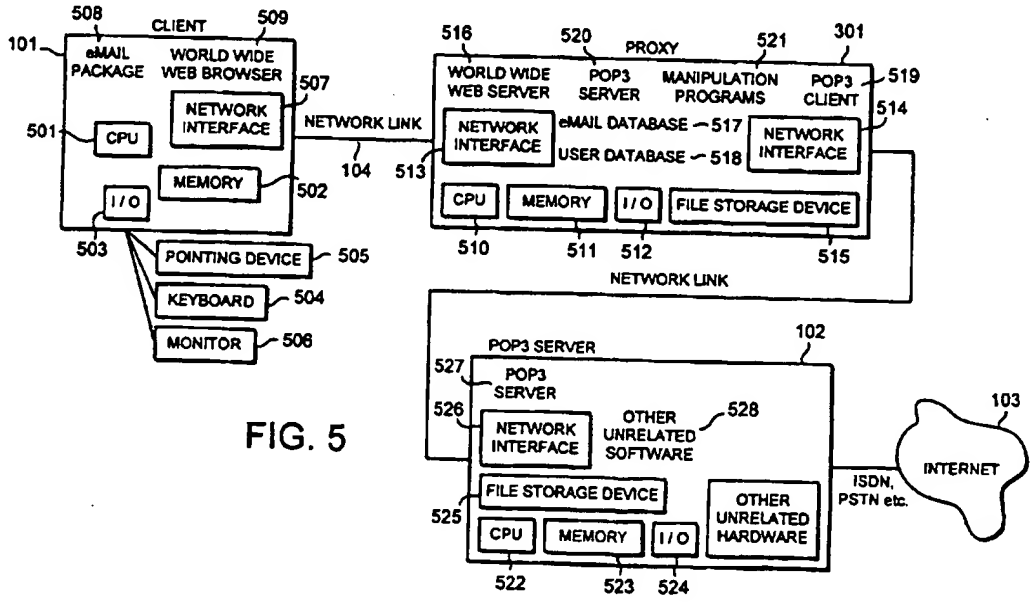

FIG. 5

GB 2 347 053 A

At least one drawing originally filed was informal and the print reproduced here is taken from a later filed formal copy.

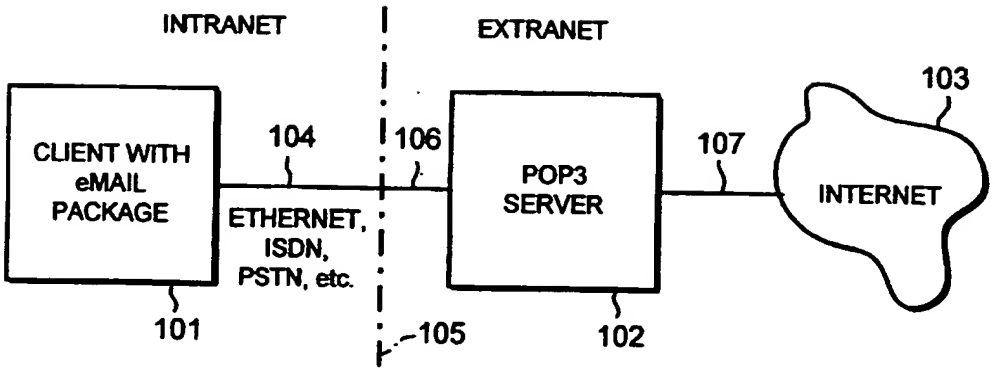
INTRANET
EXTRANET
CLIENT WITH eMAIL PACKAGE
104
ETHERNET, ISDN, PSTN, etc.
106
POP3 SERVER
107
103
INTERNET
101
105
102

FIG. 1

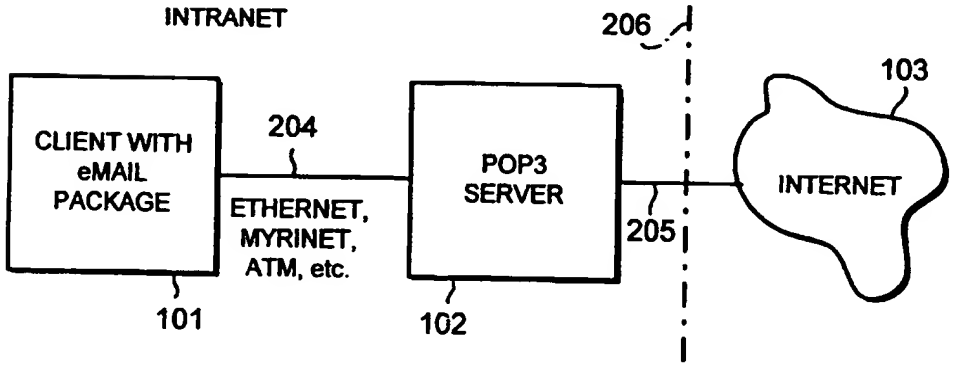
INTRANET
206
CLIENT WITH eMAIL PACKAGE
204
ETHERNET, MYRINET, ATM, etc.
POP3 SERVER
205
103
INTERNET
101
102

FIG. 2

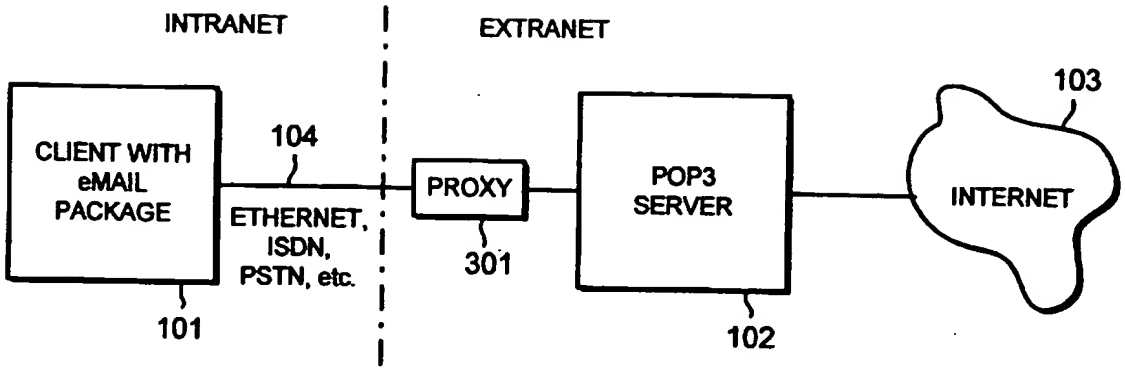
INTRANET
EXTRANET
CLIENT WITH eMAIL PACKAGE
104
ETHERNET, ISDN, PSTN, etc.
PROXY
301
POP3 SERVER
INTERNET
103
101
102

FIG. 3

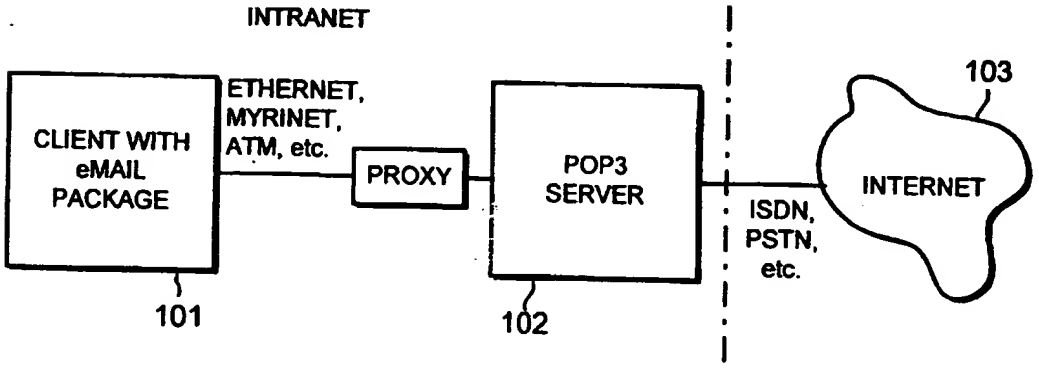
INTRANET
CLIENT WITH eMAIL PACKAGE
ETHERNET, MYRINET, ATM, etc.
PROXY
POP3 SERVER
ISDN, PSTN, etc.
INTERNET
103
101
102

FIG. 4

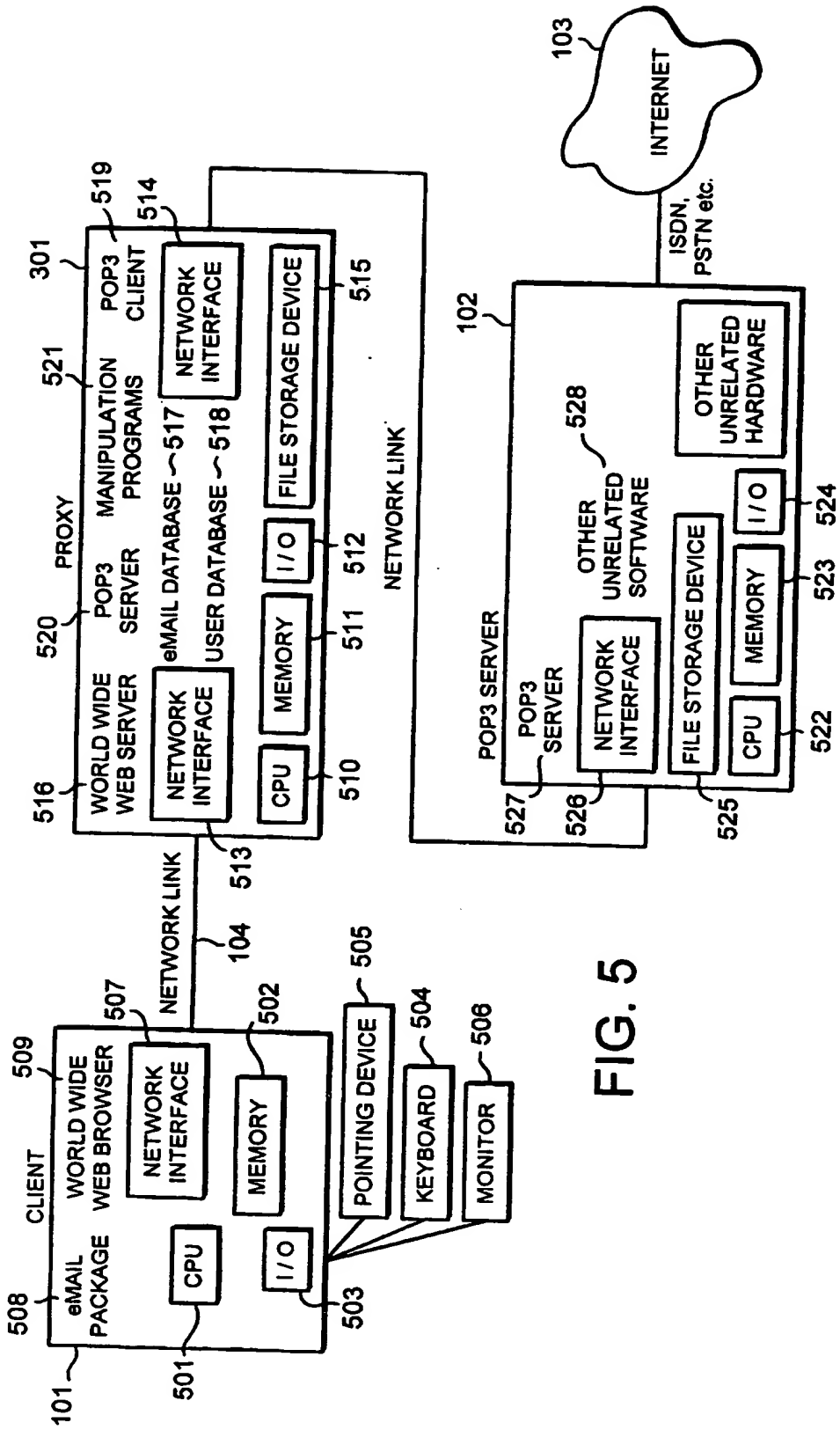
101
CLIENT
508
eMAIL PACKAGE
509
WORLD WIDE WEB BROWSER
501
CPU
503
I/O
502
MEMORY
507
NETWORK INTERFACE
NETWORK LINK
104
505
POINTING DEVICE
504
KEYBOARD
506
MONITOR
513
516
WORLD WIDE WEB SERVER
520
PROXY
301
POP3 SERVER
521
MANIPULATION PROGRAMS
519
POP3 CLIENT
514
NETWORK INTERFACE
517
eMAIL DATABASE
518
USER DATABASE
515
FILE STORAGE DEVICE
512
I/O
511
MEMORY
510
CPU
NETWORK LINK
103
INTERNET
ISDN, PSTN etc.
102
POP3 SERVER
527
POP3 SERVER
526
NETWORK INTERFACE
528
OTHER UNRELATED SOFTWARE
525
FILE STORAGE DEVICE
522
CPU
523
MEMORY
524
I/O
OTHER UNRELATED HARDWARE

FIG. 5

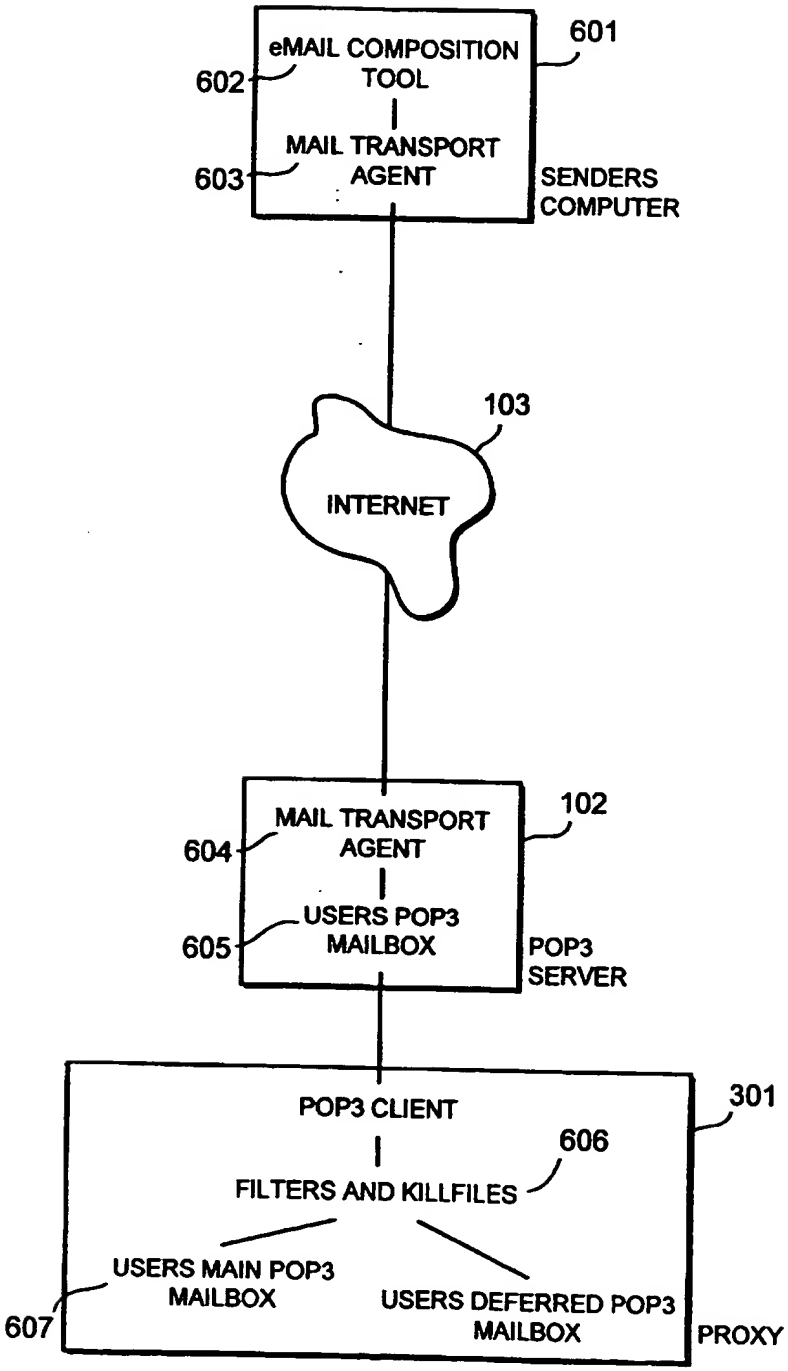
eMAIL COMPOSITION TOOL
601
602
MAIL TRANSPORT AGENT
603
SENDERS COMPUTER
103
INTERNET
MAIL TRANSPORT AGENT
102
604
USERS POP3 MAILBOX
605
POP3 SERVER
POP3 CLIENT
301
606
FILTERS AND KILLFILES
USERS MAIN POP3 MAILBOX
USERS DEFERRED POP3 MAILBOX
607
PROXY

FIG. 6

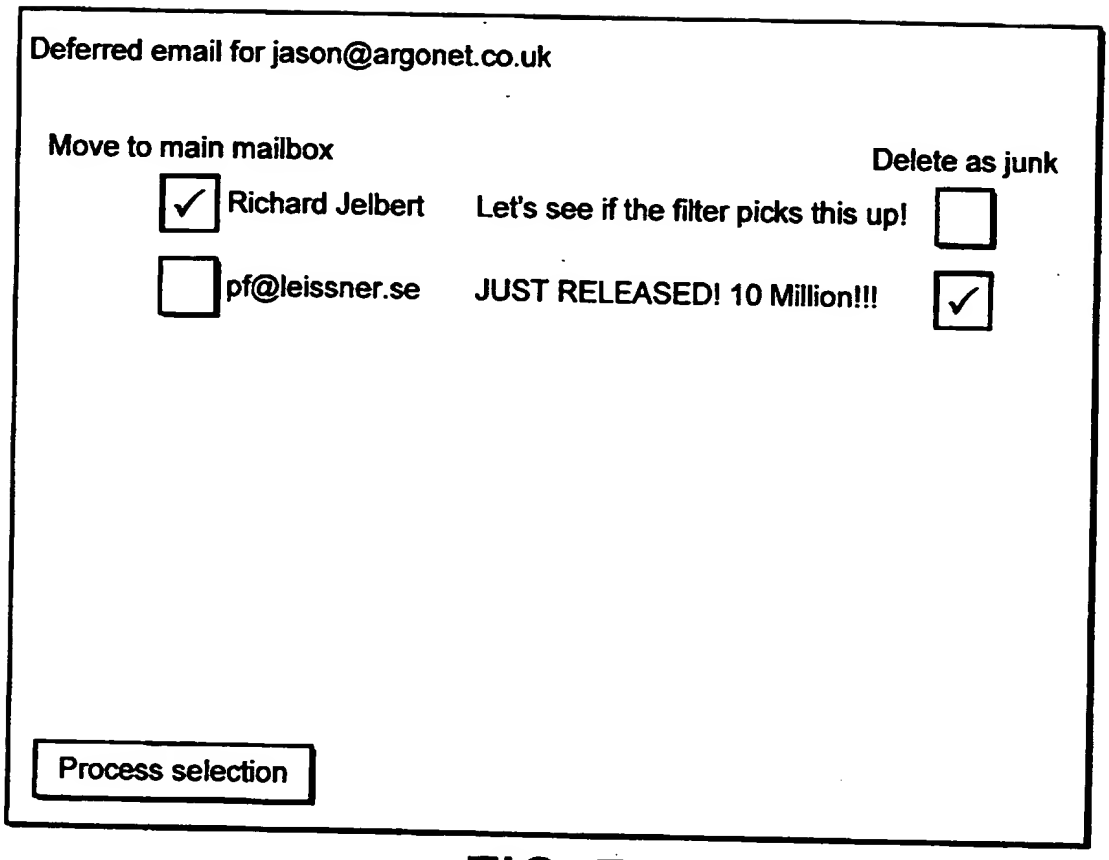
Deferred email for jason@argonet.co.uk
Move to main mailbox
Delete as junk
Richard Jelbert
Let's see if the filter picks this up!
pf@leissner.se
JUST RELEASED! 10 Million!!!
Process selection

FIG. 7

Message - ID:
Number of intended recipients who areproxy users
Date of last record update
<UserID> <killfiled|deferred|delivered to POP3 mailbox>
<UserID> <killfiled|deferred|delivered to POP3 mailbox>
<UserID> <killfiled|deferred|delivered to POP3 mailbox>

FIG. 8

UserID on proxy
UserID on main POP3 server
Password on main POP3 server
Use POP or APOP when authenticating with main POP3 server

FIG. 9

Method and Apparatus for Proxying and Filtering Electronic Mail

This invention relates to electronic mail systems where electronic mail is passed between a client and server over a network.

A piece of electronic mail, referred to by those practised in the art as email, comprises a header component and a body component. The body component contains the message which the sender wishes to deliver to the eventual recipient; this message may be a piece of plain ASCII text, or a binary (a piece of machine-readable data such as a database or spreadsheet file, or even a program designed to be executed on a particular machine architecture) suitably encoded for transmission as an email message, or a number of pieces of text and binaries encapsulated via an appropriate metric within the one whole message body. The header component, which is sometimes referred to as an "envelope", contains a number of fields; each field comprises a string of characters, and can be decomposed into a header (a piece of plain ASCII terminated by a colon), a field body, and a terminator (a carriage return followed by a linefeed).

Email messages are generally constructed according to the *Standard for the Format of ARPA Internet Text Messages* specification; this document is known to those practised in the art as RFC822, and a plaintext copy of it may be found at ftp://ftp.sunsite.doc.ic.ac.uk/rfcs/rfc822.txt . All email messages that are intended to be propagated across the Internet must conform to this specification.

The fields most commonly found within an email header are:

- The "To:" field: This field is filled in by the message sender, and contains a comma-separated list of the email addresses of the intended primary recipients. It can also point to the email address of an alias expander (see below), or to the sender if the "Bcc:" method of sending (see below) is used.
- The "Cc:" field: This field is filled in by the message sender, and contains a comma-separated list of the email addresses of the intended secondary recipients. It can also

point to the email address of an alias expander (see below).

- The "From:" field: This field is filled in by the email software running on the sender's computer, and contains the email address of the sender.
- The "Return-path:" field: This is the email address to which mail servers and routers should send a delivery error report if, for some reason, the email cannot be delivered successfully to the recipient.
- The "Envelope-to:" field: This is the email address of a single intended recipient. If this address is not listed in the "To:" or "Cc:" fields, it can be inferred that the message was sent using the "Bcc:" method (see below).
- The "Date:" field: This field is filled in by the email software running on the sender's computer, and contains the date and time at which the message was sent.
- The "Message-ID:" field: This field contains a number generated by the sender's computer, using a metric which guarantees that the number can be used to uniquely identify a given piece of email.
- The "Subject:" field: This field is filled in by the message sender, and is conventionally used to give an indication of the subject matter to which the message body content pertains. If the message is a reply to a previous message, it is conventional for the field body to begin with the string "Re:" or "Re[*n*]:" where *n* is an integer, or some string representing "Re:" or "Re[*n*]:" in different case, followed by the field body present in the message being replied to.
- The "Reply-To:" field: This field is usually, but not always, filled in by the email software running on the sender's computer. It contains the email address to which replies to the message should be sent, if for some reason it is not possible or not appropriate to reply by sending email to the address given in the "From:" field.
- The "Received:" field: This field is added to the header by a mail forwarder program, which needs to be running on computers which bridge the individual networks between the sender's and recipient's computers and thus form the complete path by which the message is propagated from sender to recipient. This field usually contains the name and Internet address of the machine on which the forwarder is running, the name and version number of the piece of software acting as the forwarder, the name

of the machine from which the forwarder received the message, the transport protocol used to transfer the message, an intermediate message ID assigned by the message transport, and the date and time at which the message was received by the forwarder.

Other headers can be inserted at the time of sending either by the sender or the sender's email program; headers which fall outside the scope of those headers defined by RFC822 and associated RFCs are given field headers which begin with the string "X-".

In addition to the "To:" and "Cc:" fields, most email clients support a pseudo-field called "Bcc:". The "Bcc:" field itself never appears in an email header as viewed by the recipient, however it may be used by the sender to instruct their email client to "Blind carbon copy" additional recipients by sending copies of the email with the "Envelope-to:" field containing the email address of each such recipient. A recipient can deduce that a message was sent to them using the "Bcc:" method if their email address, or the email address of an alias expander to which they are subscribed, does not appear in the "To:" or "Cc:" fields in the email header.

Email clients submit composed messages to a computer program known to those practised in the art as a mail transport agent, which is the program which causes a message to be propagated to its intended recipients.

It is known, and can also be inferred from the list above, that well-defined mechanisms exist for recipients of an email to reply to it; by specifying the contents of the "To:" field suitably, a user can reply to the sender of the original email, the sender of the original email and all the other recipients the sender specified, some subset of the original sender and the specified recipients, or some subset of the original sender and the specified recipients and additional people to whom the original mail was not sent. Thus, by use of the reply mechanism and the "Re:" indicator in the subject line, an ongoing discussion can evolve.

This process has expanded into the concept of distribution lists. A distribution list is generally

directed towards a particular subject matter (for example regdevs@acorn.co.uk, which dealt with technical news pertinent to developers of hardware and software for Acorn computers), thus those users who are interested in the subject matter can be "subscribed" by arrangement to the list. Their email address is then added to an alias expander, such that when an email is sent to the email address owned by the alias expander, the expander redistributes the email to the email addresses of all the list subscribers, using conventional email. Distribution lists can be managed directly by an administrator or trusted user, by a computer-executable program (such as Smartlist), or by a combination of both.

With distribution lists, particularly unmoderated ones, a subscribed user often loses interest in the subject being discussed; when the user would rather not read a message which has been sent to them by an alias expander, that message becomes electronic junk mail; the name is given as an analogy to paper junk mail, which is considered a waste of time to open or read. Another source of junk email is spam; this term is applied to email messages, often containing advertisements for products or services as the body text, which are sent to alias expanders devoted to other topics, or directly to users who have often had no prior contact with the organisation originating the spam message. If a user is subscribed to multiple distribution lists, he can often receive multiple copies of the same spam message. Spam has been described by journalists as "an obnoxious, netwide epidemic", and has even engendered lawsuits. A still further source of junk email is subscription or unsubscription requests applicable to a distribution list, which are sometimes sent by a user in error to the list's recipients rather than the list maintainer.

Version 3 of the Post Office Protocol ("POP3") is defined in two documents known to those practised in the art as RFC1939 "*Post Office Protocol - Version 3*" and RFC1957 "*Some Observations on Implementations of the Post Office Protocol (POP3)*"; copies of these documents can be found at ftp://sunsite.doc.ic.ac.uk/rfcs/rfc1939.txt and ftp://sunsite.doc.ic.ac.uk/rfcs/rfc1957.txt . POP3 is intended to permit a client to dynamically access email stored on a server in a simple fashion; the server receives incoming mail intended for a given recipient from other Internet-based servers and collects it in a defined area of

filespace (a "mailbox"), and the client (usually a workstation or thin client device) connects to the server from time to time and, by use of a small command set, is able to authenticate itself as a registered user with the server, to negotiate with the server to determine whether new mail is waiting to be collected, to make a local copy of the mail, and to delete mail from the mailbox on the server. An extension of POP3 is known as APOP, which addresses some system security issues presented by POP3 (which, in non-APOP form, requires that the User ID and associated password are transmitted as unencrypted ASCII from the client to the server) by using the user's password to encrypt a one-time unique piece of plain ASCII passed by the server to the client at connection time as a digest using the MD5 encryption algorithm, and passing this digest to the server. Details of the MD5 encryption algorithm can be found in RFC1321, a copy of which is located at ftp://sunsite.doc.ic.ac.uk/rfcs/rfc1321.txt .

The POP3 transport is designed for use in situations where the connection between the client and the server does not constitute a permanent link, or when the link between client and server is of very restricted bandwidth (ie significantly slower than 10BaseT Ethernet); thus POP3 is widely used by Internet Service Providers who provide Internet connectivity via the public switched telephone network for home users or small offices. The simplicity of the POP3 negotiation protocol also makes it very suitable for use with end-user client machines which have limited local computing power and little or no local storage, which are known to those skilled in the art as "thin" clients.

The World Wide Web has several aspects; the first aspect is a language known as the Hypertext Mark-up Language (HTML), in which documents to be made available via the World Wide Web are written. HTML documents can comprise text, graphics and interactive features, such that an element of text of a graphic can form a link to another document; the user selects a link element on a page, and the page linked to is loaded.

The second aspect of the World Wide Web is the Uniform Resource Locator (URL), and the associated Hypertext Transport Protocol (http). URL sytax is defined in the document known to those practised in the art as RFC1630, and can be found as

ftp://sunsite.doc.ic.ac.uk/rfc/rfc1630.txt . By entering a URL into an application running on a networked computer which understands how to parse URLs and perform http fetches, a user can cause the computer to request ("fetch") a local copy of an HTML document from a publicly-exported location on another computer connected to a network which can be routed to the local computer. Documents can also be specified in URLs which are local to the computer running the browser; the fetcher can then offer up the document from the computer's local storage medium if it can be found there.

The third aspect of the World Wide Web is the browser; this is an application run by a user on their local computer which is able to render documents which have been written in HTML, and which communicates with the local http fetcher such that HTML documents fetched by the fetcher are rendered by the browser.

The fourth and final aspect of the World Wide Web that is considered here is the server. This is a program run on a computer such that page fetch requests made of it by remote computers can be parsed and, if the relevant HTML document is available on the computer executing the server program and authentication conditions are satisfied, the document can be sent to the appropriate remote computer. The server supports an application interface codified as the Common Gateway Interface (CGI); this interface allows specific HTML document elements (check boxes, buttons, text areas etc) which can have their state changed by a user operating a browser viewing that document to communicate their state to a secondary application running on the machine acting as a server (and referred to by those practised in the art as a CGI script), and also allows appropriate scripts to dynamically generate customised HTML documents, which can then be served.

Formal specifications for HTML and http can be found at http://www.w3c.org/ , and details of a widely-used World Wide Web server and CGI can be found at http://www.apache.org/ .

Known systems partially address the problem of junk email by providing filters which may be applied to email by a recipient; these examine the message for a match to some condition within

their "From:" or "Subject:" fields, or within their body text. The principal problem with this approach is that junk email tends not to have a consistent set of characteristics within this scope which can readily be matched by a filter; this results in the recipient having to constantly define and refine filters to trap emails of specific character while also trying to minimise the risk of inadvertantly trapping a non-junk email.

An alternative approach is described by European Patent Application EP-A-0,813,162, where a user can determine whether or not a given message is junk and, if so, inform a mail server of the fact so that the message can be removed from the mailboxes of other users who make use of the same mailserver.

Viewed from one aspect the present invention provides apparatus for processing electronic mail, said apparatus comprising:

mail fetching logic for fetching an electronic mail message for a user from a first mail server, said apparatus interacting as a first mail client with said first mail server;

mail filtering logic for identifying at least one predetermined characteristic within said electronic mail message that is indicative of said mail message being unwanted by said user so as to identify said electronic mail message as either a wanted electronic mail message or an unwanted electronic mail message;

mail storage for storing at least wanted electronic mail messages identified by said mail filtering logic; and

mail delivery logic responsive to a mail delivery request from a second mail client for delivering wanted mail for said user from said mail storage to said second mail client, said apparatus interacting as a second mail server with said second mail client.

Transmission of the electronic mail between the client and the server may use one of several known mail transport protocol known in the prior art. Preferred embodiments of the invention uses the POP3 protocol (rather than a protocol such as IMAP, although IMAP could be employed) between the end-client and the main server, since the use of POP3 enables the invention to be added to a pre-existing system without any change having to be made to the

server in such a pre-existing system. In the treatment below the mail transport between the existing server and the filtering system, and the filtering system and the end-user client, is treated as being POP3, however it must be noted that the scope of the invention is not limited to embodiments where POP3 is used for this purpose.

Specifically, the invention may comprise a proxy system (whether as a distinct separate computer apparatus or a modular software component) which can be inserted between a POP3 server and a POP3 client, such that the system takes electronic mail in from the POP3 server, automatically filters it for junk mail according to a set of rules, and then passes the filtered mail out to the client via a second POP3 stream (thus appearing to the original client to be a POP3 server). In this configuration, no changes need to be made to the original POP3 server, and the only changes which need to be made to the POP3 client involve configuring it to receive POP3 from the proxy rather than the original server. The advantages of this system over systems known in the prior art are that the administrator of the proxy does not need to have administration privileges on the POP3 server (since, in the preferred embodiment of the invention, no modifications need to be made to it), and as filtering is performed by the proxy rather than by the clients, the filtering process requires no more local computation on the part of the clients than would be required if the proxy was not in place (this is important when the clients are of a type known to those practised in the art as "thin" devices, as such devices only have a very small amount of computing power with which to process email and perform other tasks).

Electronic mail considered by the system to be junk following automatic filtering can, optionally and dependent on the configuration set by the proxy's administrator, either be discarded or placed in a per-user "deferred" mailbox rather than being delivered by POP3; this "deferred" mailbox is an area of mass storage on the proxy where electronic mail can be stored, retrieved and presented to its original intended recipient. In the preferred embodiment of the invention, the contents of the deferred mailbox can be presented to the user for inspection via a secure World Wide Web page. Users may access their deferred mailbox via a World Wide Web browser, and using check boxes and action buttons can elect to have messages moved from the deferred box to their main POP3 box, or deleted.

Should any particular electronic mail message that passes the screening metrics be found to comprise junk on inspection by the message's intended recipient, the user can inform the server of the junk nature of the message using a World Wide Web interface. If a number of users (the number is configured by the proxy's administrator) mark a particular message as being junk, then it is automatically deleted from the system. The deferred mailbox is subject to automatic message deletion for messages which have been resident there for an administrator-configurable length of time (a week is suggested as an appropriate time interval) to prevent the deferred mailboxes becoming too large.

Embodiments of the invention will now be described, by way of example only, with reference to the accompanying drawings in which:

Figure 1 illustrates the connectivity between an email client, a POP3 server and the Internet, as typically observed between a single end-user and an Internet Service Provider;

Figure 2 illustrates the connectivity between an email client, a POP3 server and the Internet, as typically observed between a small business running a number of computers (each hosting an email client), and an Internet Service Provider;

Figure 3 indicates where in the chain of connectivity shown in Figure 1 a preferred embodiment of the invention would be installed by an Internet Service Provider;

Figure 4 indicates where in the chain of connectivity shown in Figure 2 a preferred embodiment of the invention would be installed by the system administrator of a small business' computer network;

Figure 5 illustrates portions of the computer systems shown in the above figures, and indicates the hardware present in and the software running on a preferred embodiment of the invention;

Figure 6 illustrates the principal computer programs which handle an email message in the preferred embodiment; the email message is originated at the top of the diagram, and received in a mailbox at the bottom of the diagram;

Figure 7 shows an example of what the World Wide Web interface presented to the user for examining and manually filtering the contents of his deferred mailbox may look like;

Figure 8 shows the assignment of fields within a record contained within the email message database held on the proxy, in the preferred embodiment; and

Figure 9 shows the assignment of fields within a record contained within the database of users of the proxy, in the preferred embodiment.

In the figures above, items of hardware within a computer system are differentiated from items of software running on that computer system by enclosing the character strings naming the items of hardware within a rectangular box.

Overview

The system allows email messages resident on a POP3 server to be accessed via a proxy, such that the proxy performs filtering operations on the email as it is downloaded by a user of the proxy. Optionally (according to the configuration set by the proxy's system administrator, and if the proxy is running on hardware separate from the main POP3 server), all email stored on a remote POP3 server which is intended to be received by a user of the proxy may be "harvested" by the proxy at an administrator-configurable time and cached within the proxy. The filtering operations take place remote from the end-user system, such that this end-user system does not have to devote local computing power to performing the filtering operations, and the only change in configuration required at the end-user system involves changing the information pertaining to which POP3 server the user's email client should point at. The POP3 server

requires no modification.

End-users who are able to connect to the POP3 proxy using POP or APOP and associated authentication, by virtue of them having an account on the proxy apparatus, comprise a trusted group. Each user has two areas of filespace hosted on the proxy to which they have access; a main mailbox, which can be accessed using POP3, and a deferred mailbox. If an email to a user is scored by the filters as being suspected junk email, it is moved into the user's deferred mailbox, where it can be viewed by the user via an authenticated World Wide Web interface. By using interactive elements within the World Wide Web interface, such as buttons and checkboxes, the user can inform the proxy which emails within the deferred mailbox are genuinely junk, and which should be moved into the user's main mailbox such that they can be collected via POP3. The proxy maintains a database of all the email messages it holds, and if a sufficient number of users mark a given message as junk, the message is removed from all users' mailboxes so that users who have yet to read it do not have to waste their time doing so. Email messages held within the deferred box are automatically deleted a configurable time after they have been received, so that the deferred mailboxes do not grow in size to a point where they fill the file storage system installed on the proxy to its capacity.

Thus, the group of people who use the proxy benefit from having their POP3 email automatically filtered such that junk messages are not presented for download; however, they also have a second area containing the messages which have been filtered out of their POP3 mailboxes, which they can elect to read to determine which messages are actually junk. If they find a junk message and inform the server of the nature of the message via their World Wide Web interface, the message is removed from all users' mailboxes, thus the first readers of a junk message benefit the group as a whole.

Operating Environment

Figure 1 shows a typical arrangement for a user client **101** to connect to POP3 server **102** and thence to the Internet **103**, where the dashed line **105** indicates the point at which Internet and

Intranet connect (routers, modems, modem concentrators and other equipment used to link the physical media **104**, **106** and **107**, where **104** is usually a telephone or ISDN line and **106** and **107** are Ethernet or some other fast interconnect commonly utilised between machines located at the same physical site, along which the email travels, are not shown), and where the client has no server support local to it (this is the most common configuration for an Internet Service Provider serving home users), and Figure 2 shows an alternative arrangement for the network such that the POP3 host **102** lies within an Intranet and consequently has a much higher-bandwidth link **204** (such as a 10BaseT Ethernet link) to the end-user client systems **101**. This arrangement is more typical of a small office running thin clients as end-user terminals; the slower ISDN or telephone link **205** to the Internet **103** is located at the far side of the POP3 server from the clients.

Figures 3 and 4 illustrates the location of the proxy **301** within the networks illustrated in Figures 1 and 2; Figure 5 gives a more detailed breakdown of the thin client or multiple client case illustrated in Figure 4, showing the individual computing elements (CPUs, network interfaces, memories, file storage units) present in an end-user thin client terminal, a POP3 proxy host, and a POP3 server. On the user's thin client **101**, a Central Processing Unit (CPU) **501** is connected to a memory **502** and an I/O controller **503**; the I/O controller supports a keyboard **504**, a pointing device **505** and a display system **506** such as a monitor or a domestic television. The thin client contains a network interface **507** (which may be an Ethernet adaptor, a modem designed to operate over the public switched telephone network, an ISDN modem and terminal adaptor, a cable modem, or some other network interface), and executes locally to itself an appropriate network stack, a World Wide Web browser **508** and an email package **509** which supports email reception via the POP3 protocol. The thin client has no local file storage capability, but instead makes use of file storage on a server (this could be the main POP3 server acting in another functional capacity, or a different server altogether). It links to the proxy apparatus **102** via an appropriate network connection **106**.

The proxy apparatus hardware **301** comprises a CPU **510**, a memory **511**, an I/O controller **512** and either one or two network interfaces **513**, **514** (dependent on the exact structure of the network into which the proxy is to be installed) of types described above; the I/O controller **512**

supports a file storage device **515** such as a hard disc drive, and although the I/O controller also has support for a keyboard, a pointing device and a display system such as a monitor, it is intended that these devices will not need to be permanently connected once initial configuration and installation has been performed (the intention being that all subsequent administration is performed via an authenticated World Wide Web interface). The proxy apparatus executes locally to itself an appropriate network stack or stacks, a World Wide Web server **516**, a database **517** to contain details of the nature and status of email messages stored locally on the file storage device, a database **518** to contain user mapping details, a POP3 client **519**, a POP3 server **520**, and a set of database and World Wide Web server manipulation programs **521** which embody the filtering system.

Alternatively, the proxy can be implemented purely as a set of software components on the main POP3 server **102**; the POP3 server (which comprises a CPU **522**, a memory **523**, an I/O controller **524**, one or more file storage devices **525** and one or more network interfaces **526**, and which already runs locally to itself one or more network stacks, a mail transport agent and a POP3 server **527**, as well as being likely to run computer-executable code to perform other unrelated services **528**) can in this instance have added to it a POP3 client **519** to communicate with the existing POP3 server via an internal calling and message passing mechanism, a World Wide Web server **516**, two databases **517** and **518**, a POP3 server **520**, and a set of database and World Wide Web server manipulation programs **521** which embody the filtering system.

Figure 6 provides a conceptual overview of the elements of an electronic mail system which features message interchange over the POP3 protocol and which uses the system disclosed here to provide message filtering services. A message sender's email system **601** contains a composition facility **602** that allows the sender to compose an email message, including specifying a list of recipients and a subject. This email is passed to a mail transport agent **603**, where it is sent to the addresses of the intended recipients. Often, the message is sent to a remote computer by using the Internet **103**; if an intended recipient has an address on the same computer as the sender, the Internet is not used but instead message deliver is handled by the computer which hosts the sender's and recipient's accounts. Optionally, a copy of the message is stored in the sender's filespace as resident either on the system they are sending from or some

other filesystem on a server local to their site. If the message is destined for a user who has an address handled by POP3 server **102**, the message passes across the Internet, is handled by the POP3 server's mail transport agent **604**,and is passed into the user's POP3 mailbox **605** as resident on the file storage device present on that system. If the embodiment of the filtering system in use involves the use of a separate hardware apparatus **301** as the proxy, the message is left in the mailbox **605** ready for collection when the user connects from their client to the proxy (if the proxy is not operating in harvest mode); if the elements of the filtering system are installed on the same hardware which hosts the POP3 server, the mail is collected locally. The mail is filtered using processes **606** such as those detailed in the "Automatic Mechanisms for Filtering and Scoring Email" section below and delivered either to the user's main POP3 mailbox **607** as resident on the proxy, or to the user's deferred mailbox **608.**

Operational Overview

Message Fetching, Automatic Filtering and POP3 Delivery

At some time convenient to the user, the user elects to examine his POP3 mailbox to determine whether new mail has arrived. If the embodiment of the filtering system in use involves the use of a separate hardware apparatus as the proxy, the user connects to the POP3 server running on the proxy device and authenticates himself using the recognised POP or APOP authentication mechanisms. If the proxy device is configured to filter mail on a per-connection rather than a "harvesting" basis, the proxy device observes that an authorised user has connected, and looks up in a database the appropriate address of the user's main POP3 server and the UserID and password with which the user would authenticate himself on that server. The proxy then contacts the user's main POP3 server, authenticates itself with the server using the connected user's UserID and password via the recognised POP or APOP authentication mechanisms, and uses the recognised POP3 mechanism to transfer any messages waiting in the user's mailbox to itself before terminating the POP3 session. For each new message fetched, the proxy then performs the following operations in the following order:

•The proxy checks the message headers against the criteria embodied in the global killfile

maintained by the proxy's system administrator, discarding any messages which match the criteria specified in the killfile for explicit message discarding.

•The proxy checks the message headers against the criteria embodied in the user's killfile as maintained by the user and which is to be applied to messages for which he is the intended recipient, discarding any messages which match the criteria specified in the killfile for explicit message discarding.

•The email database is checked for the existence of a record matching the "Message-ID:" field of the message; if no such record is found, the "Message-ID:" field, an integer indicating the number of intended message recipients who have addresses matching addresses of authorised users of the proxy, a field indicating the message's status (filtered as valid, filtered as junk or manually classified as junk) for the named user, and the time of receipt of the email are formed into a record and added to the database. If a database record matching the "Message-ID:" field of the message is found, the record is examined to determine whether the user has already received a copy of the message; if so, the message is discarded. If the user has not already received a copy of this message, the record is extended by adding a field indicating the message's status (filtered as valid, filtered as junk or manually classified as junk) for the named user is added, and the field representing the time of receipt of the email is updated with the time of receipt of this copy of the email.

•The proxy then enumerates the scoring filters installed (scoring filters are described in the "Automatic Mechanisms for Filtering and Scoring Email" below) and submits the message header and body text to each filter in turn. Each filter returns a positive integer; the higher the value of the integer, the closer is the match between the message submitted and what the filter considers to be junk email.

•The proxy multiplies each of the integers returned by the filters by an individual weighting (configured by the system administrator, and reflecting his confidence in the ability of each filter to reliably isolate junk email from useful email), sums the weighted integers, and compares the sum against threshold values set by the message's intended recipient and the system administrator in their configuration files. The user does not have to supply a threshold value; if the user does not supply a value, then the value set by the proxy's system administrator is used by default. If both the system administrator and the user have supplied values, the higher value is used for the comparison.

•If the sum is greater than or equal to the threshold value, the message is moved into the user's deferred mailbox. If the sum is less than the threshold value the message is copied into the user's main mailbox, from where the user can retrieve it via a POP3 transaction forming part of his current session.

If the proxy device is configured to operate in a "harvesting" mode, the proxy will, at a time configurable by the proxy's system administrator, establish POP3 connections with the main POP3 server of each authorised user of the proxy in turn, authenticate itself with the main POP3 server as the appropriate user, fetch any waiting email from the mailbox of each user in turn, close the POP3 connections, and for each message in turn, operate upon the message using the processes listed above.

If the embodiment of the filtering system in use does not involve the use of a separate hardware apparatus as the proxy, the proxy will collect messages from the main POP3 server at or very shortly after the time when the message, having been processed by the POP3 server's mail transport agent, arrives in the user's POP3 mailbox. The messages will then be operated upon in turn using the processes listed above.

Presentation of the Deferred Mailbox and Manual Junk Email Classification

The contents of a user's deferred mailbox may be operated upon by the user via a World Wide Web interface. An HTML document is constructed by a CGI script using methods known in the prior art, such that the user's World Wide Web browser presents him with a document containing information ("From:" field, "Date:" field, "Subject:" field, etc) pertaining to each message in the deferred mailbox. Each set of information pertaining to a particular message constitutes a link, such that the message headers and body text are presented in full as another HTML document (HTML mark-up being performed by another CGI script) if the link is clicked on by the user's pointing device.

Each set of information pertaining to a particular message will have two checkboxes associated with it; checkboxes are user-interaction elements that comprise part of a "fill-out form" as

known to those skilled in the art of HTML. One checkbox has the function of marking the message as junk, and the other has the function of marking the message as valid and moving it to the user's main mailbox for delivery via POP3. An example of the possible display is shown in Figure 7.

Database Information

A preferred embodiment of the invention maintains information in a database regarding the unique "Message-ID:" identifiers for email messages, the number of intended recipients of each message who are authorised users of the proxy, the UserIDs of the authorised users who have received the email messages, and whether the killfile and configuration file of each user who is an intended recipient of a given message has caused that user's copy of the message to be automatically discarded, stored in the deferred mailbox as junk, or stored as useful mail in the main mailbox for POP3 delivery. The structure of a record of this type is represented in Figure 8, and the database is indexed on the contents of the "Message-ID" field.

A second database holds records pertaining to authorised users. The proxy is able to operate in conjunction with multiple main POP3 servers, therefore each record in this database is indexed by the UserID of the user as registered with the proxy, and contains fields representing the address of the user's main POP3 server, the user's UserID on the remote POP3 server, the user's password on the remote POP3 server, and a flag to indicate whether the connection to the remote POP3 server should be authenticated using POP or APOP. The structure of a record of this type is represented in Figure 9.

Database and Mailbox Maintenance

The email message database and the users' deferred mailboxes will grow in size over time, so a periodically-operating mechanism must be put in place to prevent them growing to a capacity which will fill and attempt to overflow the file storage device used by the proxy. Such a mechanism would in a preferred embodiment operate daily, and be scheduled to take place at a

time when computer activity is predicted by the proxy system administrator to be low (activating the maintenance process at a time of 03:20 am is suggested).

The email message database is examined, starting from the first record, and continuing sequentially so that all records are examined in the process. The ratio of the number of users who have specified the mail item as being junk to the number of users who have not specified the message as being junk and the absolute number of users who have specified the mail item as being junk is enumerated, and if this exceeds a ratio or absolute value set by the proxy system administrator, the message is deleted from the deferred mailboxes of all the users indicated in the record as having received it.

The "Date of last update" field is compared with the current calendar date, and if the difference in dates between the date of update and the current date is equal to a suitable threshold set by the proxy's system administrator, the record is deleted. A system administrator knowledgeable of the typical propagation times of email through the various mechanisms present in the Internet will be aware of the fact that two copies of the same message can arrive with a significant time delay between them, and it is suggested that the threshold value for deletion of a record is set to seven days after the record was last updated.

The deferred mailboxes require a different approach; users may be unable to examine their deferred email for a number of weeks owing to holiday, illness, business trips etc. Hence it would be inappropriate except in extreme circumstances for the contents of these mailboxes to be pruned automatically; instead, it is suggested that a periodically-operating mechanism (again, operating daily in the preferred embodiment) would enumerate the authorised proxy users, determine the size of each user's deferred mailbox, and construct a digest in the form of a World Wide Web document which would then be placed in the World Wide Web document area accessible only by the proxy's system administrator. If the system administrator saw that a particular user's deferred mailbox was becoming excessively large, he could then take appropriate action in accordance with his organisation's policy.

Automatic Mechanisms for Filtering and Scoring Email

Administrator and User Killfiles

A killfile, is usually applied to USENET newsreaders rather than email systems. It comprises a list of email addresses and / or keywords found in header lines or body text, and information on how mail messages that match them should be dealt with. Following the model used in the configuration files of the Apache World Wide Web server to determine who is allowed access to a site, the email proxy can be configured either to stop all mail being passed through except that from specified users or with specified words in a header line or the body text, or to allow through all email except that from specified users or with specified words in a header line or the body text, or to copy emails from specified users or with specified words in a header line or the body text to a human administrator for examination. Such entries could therefore check for the user name of the sender, the domain the email was sent from, and the type of email client software the sender used (some clients are more suited to producing spam than others)

The effects of the Administrator killfile should be treated as global, ie any entry in the Administrator killfile is applied to any item of email passing into the proxy from the server; User killfiles should have a scope of application limited to email addressed to a specific user. Killfiles stored as part of the proxy apparatus may be made available to their respective owners by a file-export mechanism such as authenticated ftp or NFS, or the killfile may be constructed and maintained on a user's workstation and sent to the proxy by email. An example of a killfile can be found in Table 1.

```
#Open policy; allow mail from everywhere, then filter
allow from all

# Lock out mail sourced from AOL
deny from *@aol.com

# ... but allow mail from Scott Adams (Dilbert) by exception
allow from scottadams@aol.com
allow from dogbert@aol.com

# Lock out mail from addresses of known spammers
deny from qoy84@prodigy.com
deny from pf@leissner.se

# Lock out messages with multple dollar signs etc in their subject lines
deny subject "$$"
deny subject "!!"
deny subject "MONEY FAST"
deny subject "XXX"
deny subject "LOSE WEIGHT"

# Lock out erroneous "subscribe" and "unsubscribe" messages on mailing lists
deny to "regdevs@acom.co.uk" AND body "unsubscribe"
```

Table 1: An example of a simple user killfile

5 Message Scoring

If an email message is not explicitly barred by killfiles installed by the system administrator or the recipient of the mail, there are a numerous ways in which it may be filtered by a computer program to determine whether or not it is junk or spam email. Autonomous filters are sometimes used, however the methods outlined below can be treated as part of a modular structure in which, once a test is performed, the result of the test can be treated as a numerical value; this value can then be added to a running total relating to the message, such that if the total after the tests have been performed is greater than or equal to a threshold value set either by the user or the system administrator, the message can either be forwarded to a trusted user to determine whether or not it genuinely is junk mail, or it can be discarded by the computer subject to configuration.

The following message scoring metrics by no means comprise the whole collection of tests that may be performed by the system; rather they serve as a useful starting point upon which to build a message-scoring suite of tests. In addition to these tests, for example, string matching and related-string matching techniques such as those used by the freeware USENET newsreader "slrn" (source code for which is available on ftp://sunsite.doc.ic.ac.uk/) may be converted into scoring programs, and particularly advanced scoring programs may employ neural network techniques to identify and weight messages against known patterns which tend to occur in junk email, particularly spam.

For the purposes of example, consider the message headers of two pieces of spam, reproduced as Tables 2 and 3.

Return-path: <qoy@prodigy.com>
Envelope-to: Jason@argonet.co.uk
Delivery-date: Mon, 13 Jul 1998 03:13:43 +0100
Received: from (maill.noc.netcom.net) [204.31.1.150] by golden.argonet.co.uk with esmtp (Exim 1.82 #2) id OyvY7a-00054F-00; Mon, 13 Jul 1998 03:13:42 +0100
Received: from ariesresearch.com (mail.ariesresearch.com [206.216.212.201] by maill.noc.netcom.net (8.8.7/8.8.5) with SMTP id TAAOI 164; Sun, 12 Jul 1998 19:02:25 -0700 (PDT)
From: qoy84@prodigy.com
Received: from IBM by ariesresearch.com (SMI-8.6/SMI-SVR4) id TAA06893; Sun, 12 Jul 1998

19:03:09 -0700
Date: Sun, 12 Jul 1998 19:03:09 -0700
To: qoy84@prodigy.com
Comments: Authenticated sender is <qoy84@prodigy.com>
Subject: 57 Million Email Addresses = $99
Message-ID: <1998071221340AA33860@pimaia7y.ari.com>
Status:
X-Mozilia-Status: 2001

Table 2: An example of a mail header attached to a spam email

Return-path: <pf@leissner.se>
Envelope-to: jason@argonet.co.uk
Delivery-date: Mon, 17 Aug 1998 22:49:34 +0100
Received: from (box.argonet.co.uk) [194.200.2.1] by golden.argonet.co.uk with smtp (Exim 1.82 #2) id Oz8X9i-0002xY-00; Mon, 17 Aug 1998 22:49:34 +0100
Received: from (golden.argonet.co.uk) [191.131.104.13] by box.argonet.co.uk with smtp (Exim 1.81 #8) id Oz8X9h-0006Pk-00; Mon, 17 Aug 1998 23:49:33 +0200
Received: from (ns2.daio-paper.co.jp) [210.151.233.197] by golden.argonet.co.uk with esmtp (Exim 1.82 #2) id Oz8X9d-0002xS-00, Mon, 17 Aug 1998 22:49:29 +0100 **Received:** from default by ns2.daio-paper.co.jp (8.8.5+2.7Wbeta5/3.3W9-NEC) id GAA04245; Tue, 18 Aug 1998 06:46:35 +0900 (JST)
Date: Tue, 18 Aug 1998 06:46:35 +0900 (JST)
From: pf@leissner.se

Received: from login-01224.roverdigger.net (mail.roverdigger3.net[195.75.899.454]) byroverdigger.net (8.8.5/8.7.3) with SMTP id XAA06218 for user1244@roverdigger.net; Tue, 18 August 1998 04:19:24 -0700 (EDT)
To: pf@leissner.se
Subject: JUST RELEASED! 10 Million!!!
X-PMFLAGS: 225549798.233
X-UIDL: 15424665-288569.564.747
Comments: Authenticated Sender is <user1224@roverdigger.net>
Message-ID: 01658742211308922@g-hipkernia.com
Status:
X-Mozilla-Status: 2001

Table 3: A second example of a mail header attached to a spam email

Now apply the following scoring mechanisms:

Header Integrity

Examine the header for the ordering of "Received:" fields; it can clearly be seen in Table 2 that, starting from the last-printed "Received:" field and working up to the first-printed "Received:" field within the header, that the message followed a path from IBM to Aries Research to Netcom and finally to Argonet. However, note that the "From:" field is inserted in such a place within the header that the flow of "Received:" fields is interrupted; normally, all the "Received:" fields would form a single contiguous block within the header. This practice of dividing the "Received:" fields is indicative of spam email. Thus the location of the "From:" field within this message header suggests a high probabililty that the message is a piece of spam, and it would be scored accordingly. Also, in Table 3, it can be seen by one skilled in the art that the string of numbers returned as the IP address of mail.roverdigger3.net (195.75.899.454) is clearly outside the range in which IP addresses are allocated; in an integrity check which verifies the existence of each host in the "Received:" chain using a protocol such as ICMP, this message would be scored as highly suspect by returning a large integer.

Can the proxy Reply to the Message Sender?

Junk email, particularly spam, often has its headers modified by the sender so that the apparent sender has a ficticious email address. This is generally done so that messages from recipients complaining about the spam do not consume resource on the sender's server. Verifying the validity of a sender's email address is therefore another valuable method of determining whether or not a message is spam.

The validity of the sender's apparent email address can be verified thus:

- Extract the "Reply-To:" field, if it is present, from the message header; extract the "From:" field if the "Reply-To:" field is not present.
- Extract the domain component (the string following the "@") from this field.

- Look up the MX records in the domain name service for that domain, and enumerate the machines in that domain which are marked as having email forwardable to them.
- Query the first host in this list for the existence of the user (the user's ID is the component of the email address preceding the "@").
- If the user is not found, send the user-existence query above to the next host on the list.
- If the user is not found by sending this query to any of the machines in that domain which can receive mail, compose a test message and perform a "send preparation" negotiation with the first machine on the list.
- If the "send preparation" negotiation is refused, this indicates that the user's ID is not an alias for an expander, and is therefore unknown at the site. Therefore the message can be considered, with a high degree of confidence, to be spam and appropriate scoring can be applied.

Common Messages received by Multiple Users

A record of the message ID of each message received by each user is maintained within a database stored on the proxy; the database is structured such that it is indexed by message ID, and the contents of each record includes a field enumerating the authorised users of the proxy who have received a copy of the same message. If a message is received by a large number of users, it will either be a legitimate circular to users of a trusted group which includes the user of the proxy, or junk.

The Sender's Host

Utilising prior art, the portion of the sender's email address following the "@" can be checked against a regularly updated list of known addresses from which spam and other junk email is known to originate. Such a list, in this instance known as the "Mail Abuse Protection System Realtime Blackhole List (MAPS RBL)" is maintained at http://maps.vix.com/ .

Subject Field and Message Body Content

In some instances, the nature of the contents of the subject field or message body text can reliably indicate whether a given email is a junk message; for example, distribution list subscription and unsubscription requests usually contain a one-line body text starting with the word "subscribe" or "unsubscribe". Similarly, spam messages tend to contain an excess of exclamation marks (a contiguous block of more than two exclamation marks is not uncommon), dollar signs, and particular phrases. This method of filtering is prior art, and operates with the limitations that junk email usually does not have a unique characteristic in these areas.

Each scoring metric, when applied to an incoming email message, can return a positive integer indicating the degree to which the metric believes the message to be junk. Once all installed filters have been passed the message for scoring, the individual scores from each metric are weighted according to the confidence the proxy administrator has in the ability of each metric to distinguish junk email from useful email, and summed. The sum is compared to threshold values defined by the proxy administrator and the user; if the sum is greater than or equal to the threshold value, the message can configurably be deleted from all users' mailboxes, or moved to all users' deferred mailboxes.

Security

Those skilled in the art will recognise that World Wide Web servers such as the one used in the preferred embodiment to present deferred mailboxes can be made secure in three significant ways; either by configuring the server to present UserID - password challenge-response authentication request to a user to verify his identity before allowing him access to the CGI-constructed document showing the contents of his deferred mailbox, or by requiring that the server authenticate and serve the documents using the secure http protocol (https), or both. Thus direct user requests specifying the explicit deletion of messages and / or the manual scoring of a message as junk would be very difficult to forge successfully.

As the UserID, password and access method for a given user to access a main POP3 server would be held in a database and managed by scripts only accessible to the proxy's system administrator (and again, securable by challenge-response and / or https), the system is as trustworthy in this respect as the proxy's human system administrator. If the proxy forms part of an Intranet and the main POP3 server is part of the full Internet, the ability of the proxy's POP3 client to perform APOP may in fact be a security enhancement if the end-user client was only able to perform ordinary POP authentication.

Conclusion

It will be appreciated and understood that the systems described above significantly enhance a conventional POP3 electronic mail system by providing a proxying system which incorporates both automatic and manual filtering mechanisms to reduce the quantity of junk email presented to users by removing from user mailboxes messages which have been classed as junk, and which, if locally located and configured, can reduce the time spent by users waiting for their email to download over a slow link to a remote main POP3 server.

Further, one skilled in the art will recognise that various modifications and alterations may be made in the preferred embodiment disclosed herein without departing from the scope of the invention. Accordingly, the scope of the invention is not to be limited to the particular invention embodiments disclosed above, but should be formally defined only by the claims set forth below and equivalents thereof.

The processes described above above may be performed by a computer program running on a computer in the embodiment described. Such a computer program can be recorded on a recording medium (for example a magnetic disc or tape, an optical disc or an electronic memory device such as a ROM) in a way well known to those skilled in the art. When a suitable reading device (such as a magnetic or optical disc drive) reads the recording medium, a signal is produced which causes a computer to perform the processes described.

At least preferred embodiments of the invention provide:

an apparatus, method, system and computer program to perform the filtering which can be added into a pre-existing email client-server system without necessarily requiring modification of or addition to the programs running on the mail server or imposing additional processing load on the client. One aspect is a mail proxy apparatus, which may be an additional computer apparatus (containing, for example, a CPU, a memory, a file storage system and one or more network interfaces, and running computer-readable code which implements a POP3 client, a POP3 server, a World Wide Web server and a database, and additional scripts to manipulate the database and the World Wide Web server) which is added to the network containing the pre-existing client and pre-existing server and which appears to the pre-existing server as the client, and which appears to the pre-existing client's email system as a mail server and secure World Wide Web server;

a set of metrics to be codified into computer-readable form for utilising a computer apparatus (comprising a CPU, a memory, a file storage system and one or more network interfaces) to filter email messages for junk and spam content by analysis of the message headers and body text of said emails;

the integration of the above metrics into a suite which can be used by a computer apparatus (comprising a CPU, a memory, a file storage system and one or more network interfaces) to automatically "score" emails in a manner relating to the likelihood of their being junk emails, and act appropriately upon the score depending upon thresholds and options set by the system administrator and the intended recipient;

a deferred electronic mail system using a CPU, a memory, and a filestorage mechanism to provide a secure World Wide Web-based presentation mechanism such that an intended recipient is able to examine and classify an email message. This aspect also includes a presentation prevention mechanism that operates to prevent messages in a given class from

being presented to other users of the proxy apparatus;

the ability of the POP3 proxy to harvest all the email for all its configured users as contained in a remote POP3 server and maintain it locally to itself for serving to its users, in case the physical link to the remote server is severed;

a computer program having computer readable code embodied in a computer usable storage medium, and which implements a POP3 mail proxy service. This code may be executed on an existing POP3 mail server, provided that the POP3 mail server is also running a World Wide Web server, such that the filtering process occurs transparently to the already-running POP3 server program, such that the client communicates with the POP3 proxy program, in which the screening processes are carried out, and which in turn communicates locally with the pre-existing POP3 mail service;

that the POP3 server component of the proxy maintains two email repositories per user; one contains email to be delivered at the user's request via POP3, and the other contains "deferred" messages (which are messages which the screening system scores as being junk) which can be accessed by using a World Wide Web browser to access an HTML document built by a CGI script, such that the user can read email messages, mark checkboxes to indicate which messages are genuinely junk, mark checkboxes to indicate which messages in this mailbox should be transferred to the POP3 mailbox, and mark checkboxes to indicate which messages in this mailbox should be discarded; and

a computer program having computer readable code embodied in a computer usable storage medium. This code, when executed on a computer, causes a computer to provide services to a recipient. If the user is the proxy's administrator, and depending upon the program configuration, the services comprise all the services available to an unprivileged user plus the ability to add to, modify or configure the metrics used for screening of incoming messages for junk content, the ability to set thresholds at or above which messages are automatically considered to be junk, the ability to specify for all users how messages considered to be junk are

to be dealt with automatically (whether they are to be moved to a deferred mailbox or discarded), the ability to specify or modify a killfile against which all messages incoming to the proxy are screened prior to being screened for subject or body content, the ability to specify the mode of operation (proxy-per-connection or harvest) mode of the computer executing the code, the ability to determine the time after which a message enters a deferred mailbox it is discarded, and the ability to authorise new users so that they may use the proxy, and remove authorisation from existing users of the proxy.

If the user is an ordinary user who uses the filtering service, the service comprises POP3 email filtered to remove junk content, World Wide Web-browsable digests of messages which have been sent addressed to the user but classified by filters as likely to be junk, the ability to specify how messages to that user are dealt with once classified by filters as likely to be junk, the ability to specify and modify a killfile against which messages to that user are screened prior to being screened for subject or body content, and the ability to flag a received email as being junk such that the program will remove instances of that email from other users' mailboxes subject to the metrics imposed by the administrator.

Various aspects of at least preferred embodiments of the invention are set out in the following clauses:

Clause 1. A computer controlled method for processing electronic mail (email) comprising the steps of:

(a) extracting email from a plurality of known email servers into a separate computer apparatus (comprising a central processing unit, a memory, a file storage mechanism and one or more network interfaces) or a software system resident and executing on the known email server;

(b) employing various computer controlled methods to determine whether a given email message is likely to constitute junk;

(c) delivering the characterised mail to either the principal mailbox of the intended recipient or, if the characterisation of the message indicates that it is likely to constitute junk, to a deferred mailbox for that recipient;

(d) providing an interactive mechanism for a user to access and examine individual messages in his deferred mailbox;

(e) providing an interactive mechanism for a user to manually classify a message within his deferred mailbox as junk email;

(f) preventing presentation of messages automatically or manually classified as junk email to other users of the apparatus or system.

Clause 2. The computer controlled methods of clause 1, part (b), where a message may be characterised by:

(a) matching header and / or body text in each message against criteria set in particular configuration files ("killfiles") by the system administrator and / or the intended

recipient of the message, and either deleting the message or passing it to step (b) below dependent upon a match;

(b) passing each message which remains undeleted by the processes in step (a) above to a set of scoring metrics, to obtain a characterisation of each message indicating the likelihood of the message comprising junk;

(c) recording specific characteristics of each message in a database such that records of receipt of multiple copies of the same message can be used by the scoring metrics in step (b).

Clause 3. The structuring of the scoring metrics in clause 2, part (b) to comprise a modular and extensible suite for message scoring, such that each metric returns a numerical result indicating the likelihood that a given message comprises junk.

Clause 4. The computer performed multiplicative weighting applied to the numerical result returned by each scoring metric in clause 3, according to a file generated and maintained by the system administrator and which reflects his confidence in each metric to reliably isolate junk email from useful email, and the summing of the weighted results to produce a single characterisation metric for each message.

Clause 5. A computer controlled scoring metric forming part of the suite in clause 2, part (b), that determines whether a message is likely to constitute spam by decomposing the message header and checking:

(a) the validity of all IP addresses in the header;

(b) whether the "Received:" fields constitute a contiguous block of fields or whether they are disjoint;

(c) whether each "Received:" field relates appropriately to its immediately neighbouring "Received:" fields (ie whether each field indicates receipt of the message from the server which added the "Received:" field immediately below it).

Clause 6. A computer controlled scoring metric forming part of the suite in clause 2, part (b), that determines whether a message is likely to constitute spam by testing the validity of the email address of the apparent sender.

Clause 7. The computer controlled method of testing the validity of an email address to be used by the metric of clause 6, and comprising:

(a) extraction of the domain component of the email address of the apparent sender;

(b) lookup of the MX records in the domain name service for that domain;

(c) enumeration of the machines in that domain which are marked as having email forwardable to them;

(d) querying in turn of each host in the enumerated list for the existence of a user with username equal to the user component of the email address of the apparent sender, until either the sender is found or all hosts in the list have been queried;

(e) if all hosts in the list have been queried and none of them have confirmed existence of the apparent message sender as a user, composing a test message to the apparent sender and performing a "send preparation" negotiation with the first host in the list.

Clause 8. The allocation of two mailboxes to each user, where messages considered to be useful are delivered to one mailbox (denoted as the "main" mailbox) and messages that are suspected to comprise junk, following computer-performed classification by the computer controlled systems of clause 1 part (b) are delivered to the other mailbox (denoted as the "deferred" mailbox), as disclosed in clause 1 part (c).

Clause 9. The computer controlled method of presentation of the contents of the deferred mailbox to its owning user as disclosed in clause 1 part (d), such that each message therein may be examined and optionally classified by the user as junk.

Clause 10. The computer controlled method of notification by which the server hosting the deferred mailbox of clause 9 may be informed by the user that a message within his deferred mailbox constitutes junk.

Clause 11. The use of a World Wide Web interface to present a computer controlled interactive digest of email messages contained within the deferred mailbox as in clause 9, and to implement the notification mechanism of clause 10 by encoding notification details within a URL to be passed to a CGI script executing on the apparatus of clause 1 part (a).

Clause 12. The computer controlled method of clause 1 part (f) of deleting a message manually classified as junk according to the computer controlled methods of clauses 9, 10 and 11.

Clause 13. A computer controlled method for automatically connecting to and downloading all pending email from a plurality of remote mail servers for all registered users of an apparatus (comprising a central processor unit, a memory, a file storage mechanism and one or more network interfaces), and filing the email according to intended recipient in appropriate mailboxes stored on the apparatus in a non-interactive batch process, such that the apparatus functions as an email proxy server.

Clause 14. A computer controlled method of indexing received email for the purpose of determining its likelihood to constitute junk, by storing salient properties of each message in a database as the message is received for use by the suite of scoring metrics disclosed in clauses 3, 5, 6 and 7.

Clause 15. An electronic mail (email) and World Wide Web system having a central processor unit, a memory, a file storage mechanism and one or more network interfaces, said system comprising:

(a) an email client mechanism and an email server mechanism, these mechanisms functioning in concert to provide an email proxy service;

(b) a set of mechanisms whereby a given email message may be examined automatically to determine whether it should be classified as likely to constitute junk;

(c) a message filing mechanism such that messages classified as likely to constitute junk and destined for a particular user are filed separately from messages classified as useful and destined for that user;

(d) a World Wide Web presentation mechanism configured to interactively present email messages classified as likely to constitute junk for inspection by their intended recipient;

(e) a classification mechanism configured to allow the intended recipient of a message to classify said message;

(f) a presentation prevention mechanism configured to prevent presentation of messages formally classified as junk to registered users of the system.

Clause 16. The system of clause 15, whereby the classification mechanism of part (e) is further configured to notify the system of manual junk email classification by submitting a URL to a CGI script.

Clause 17. The system of clause 15, whereby the classification mechanism of clause 16 includes an identifying characteristic of the junk email message.

Clause 18. The system of clause 15, whereby the mechanisms of examining email include implementations of the methods described in clauses 2, 3, 4, 5, 6 and 7, and subsequent recording of the results of examination according to clause 14.

Clause 19. The system of clause 15, whereby the presentation prevention mechanism further comprises a deletion mechanism configured to delete all instances of a message having an identifying characteristic passed to it by the method of clause 16 from all the deferred mailboxes stored on the system.

Clause 20. An electronic mail (email) apparatus configured to gather, process and proxy serve electronic mail messages, said apparatus having a central processor unit, a memory, a file storage mechanism and one or more network interfaces, said apparatus comprising a message classifying, sorting and filing mechanism and a presentation prevention mechanism configured

to prevent presentation of an email message to one or more registered users of the apparatus.

Clause 21. A signal for causing an electronic mail (email) apparatus to process electronic mail messages, said apparatus having a central processor unit, a memory, a file storage mechanism and one or more network interfaces, the signal causing the apparatus to implement a message sorting and filing mechanism and a presentation mechanism configured to prevent presentation of an email message to one or more registered users of the apparatus.

Clause 22. A method of storing data on a recording medium, the method comprising storing data representative of a signal, that causes an electronic mail (email) apparatus to gather, process and proxy serve electronic mail messages, said apparatus having a central processor unit, a memory, a file storage mechanism and one or more network interfaces; the signal causing the apparatus to implement a message sorting and filtering mechanism and a presentation mechanism configured to prevent presentation of an email message to one or more registered users of the apparatus.

Clause 23. The email apparatus of clause 20, the signal of clause 21 or the method of clause 22 whereby said presentation prevention mechanism further comprises a World Wide Web server and CGI script set configured to receive a URL.

Clause 24. The email apparatus, signal or method of clause 23 whereby said URL includes an identifying characteristic and said presentation prevention mechanism further comprises an email deletion system configured to dispose of said email message having said identifying characteristic.

Clause 25. The email apparatus, signal or method of clause 24 whereby said email deletion mechanism further comprises an email removal mechanism configured to scan a mailbox to dispose of said email message.

Clause 26. A computer program product comprising:

(a) a computer usable storage medium having computer readable code embodied therein for causing a computer to gather, process and proxy serve electronic mail messages, said computer readable code comprising:

(b) computer readable code devices to cause said computer to gather, classify, sort, file and present email messages, and to effect a presentation prevention mechanism to prevent presentation of an email message to registered users of the computer.

Clause 27. The computer program product of clause 26, whereby said classification mechanisms comprise computer readable code devices configured to cause said computer to implement message scoring metrics disclosed in clauses 3, 5, 6 and 7, and subsequent storage of the results of classification according to clause14.

Clause 28. The computer program product of clause 26, whereby said presentation prevention mechanism further comprises computer readable code devices configured to enable a computer to receive a URL containing an identifying characteristic and effect an email deletion mechanism configured to dispose of said email message having said identifying characteristic.

Clause 29. The computer program product of clause 26 whereby said email deletion mechanism further comprises computer readable code devices to cause said computer to effect an email removal mechanism configured to scan a mailbox to dispose of said email message.

CLAIMS

1. Apparatus for processing electronic mail, said apparatus comprising:

mail fetching logic for fetching an electronic mail message for a user from a first mail server, said apparatus interacting as a first mail client with said first mail server;

mail filtering logic for identifying at least one predetermined characteristic within said electronic mail message that is indicative of said mail message being unwanted by said user so as to identify said electronic mail message as either a wanted electronic mail message or an unwanted electronic mail message;

mail storage for storing at least wanted electronic mail messages identified by said mail filtering logic; and

mail delivery logic responsive to a mail delivery request from a second mail client for delivering wanted mail for said user from said mail storage to said second mail client, said apparatus interacting as a second mail server with said second mail client.

2. Apparatus as claimed in claim 1, wherein said mail filtering logic identifies a plurality of predetermined characteristics within an electronic mail message to derive a score value associated with said electronic mail message, said electronic mail message being classified as an unwanted electronic mail message by comparing said score value with a threshold score value.

3. Apparatus as claimed in any one of claims 1 or 2, wherein said plurality of predetermined characteristics include one or more of:

(i) said electronic mail message has a sender identifier matching one or more known senders of unwanted electronic mail messages;

(ii) said electronic mail message has a subject identifier or message text including text matching one or more known texts indicative of unwanted electronic mail messages;

(iii) said electronic mail message has a header with a format characteristic matching one or more known format characteristics indicative of unwanted electronic mail messages;

(iv) said electronic mail message includes a message identifier matching a message identifier of electronic mail messages sent to other users and held within said mail storage

indicating that the same electronic mail message has been sent to multiple users; and

(v) said electronic mail message has a reply address identifier that may be validly used to send a send a reply to said electronic mail message.

4. Apparatus as claimed in any one of claims 2 and 3, wherein said mail filtering logic applies a predetermined weighting to each of said predetermined characteristics to derive said score value.

5. Apparatus as claimed in any one of the preceding claims, wherein unwanted electronic mail messages are also stored within said mail storage.

6. Apparatus as claimed in claim 5, comprising unwanted mail delivery logic responsive to an unwanted mail request from a user for delivering to said user unwanted electronic mail messages held within said mail storage for said user.

7. Apparatus as claimed in claim 6, wherein said unwanted mail request is an WWW page request from said user and said unwanted electronic mail messages are returned to said user as WWW pages.

8. Apparatus as calimed in any one of claims 6 and 7, wherein said mail filtering logic is responsive to an unwanted mail confirmation signal from a user confirming that an electronic mail message is an unwanted electronic mail message to modify said at least one predetermined characteristic such that other instances of said electronic mail message received by other users are also confirmed as unwanted electronic mail messages.

9. Apparatus as claimed in any one of the preceding claims, wherein said apparatus is physically remote from at least one of said first mail server and said second mail client.

10. Apparatus as claimed in any one of the preceding claims, wherein exchange of mail messages uses the POP3 protocol.

11. Apparatus as claimed in claim 10, wherein said second mail client points to said apparatus as its POP3 mail server.

12. Apparatus as claimed in any one of claims 10 and 11, wherein said apparatus points to said first mail server as a POP3 mail server for said user.

13. Apparatus as claimed in any one of the preceding claims, wherein said mail fetching logic is triggered to fetch any electronic mail messages for said user from said first mail server by said mail delivery request.

14. Apparatus as claimed in any one of claims 1 to 12, wherein said mail fetching logic is periodically triggered to fetch any electronic mail messages for said user from said first mail server independently of any mail delivery request.

15. Apparatus as claimed in claim 5, wherein unwanted mail storage logic operates to delete unwanted electronic mail messages form said mail storage in accordance with predetermined parameters in order to recover storage capacity within said mail storage being used by said unwanted electronic mail messages.

16. Apparatus as claimed in any one of the preceding claims, wherein said mail filtering logic uses at least one predetermined characteristic defined by said user.

17. Apparatus as claimed in claim 16, wherein said user defines said at least one predetermined characteristic via a WWW browser.

18. Apparatus as claimed in any one of the preceding claims, wherein said mail filtering logic uses at least one predetermined characteristic defined by a system administrator.

19. A method of processing electronic mail, said method comprising the steps of:

fetching an electronic mail message for a user from a first mail server, said fetching being performed as if a first mail client is interacting with said first mail server;

identifying at least one predetermined characteristic within said electronic mail message that is indicative of said mail message being unwanted by said user so as to identify said electronic mail message as either a wanted electronic mail message or an unwanted electronic mail message;

storing at least wanted electronic mail messages identified by said mail filtering logic; and

in response to a mail delivery request from a second mail client, delivering wanted mail for said user from said stored mail to said second mail client, said delivery being performed as if a second mail server is interacting with said second mail client.

20. Apparatus for performing as an electronic mail message client, said apparatus comprising:

a mail delivery request generator for generating a mail delivery request to a mail server having the form of the apparatus as claimed in any one of claims 1 to 19.

21. Apparatus as claimed in claim 20, comprising means for generating unwanted mail confirmation signals to confirm to said apparatus as claimed in any one of claims 1 to 19 that an electronic mail message is an unwanted electronic mail message.

22. A computer program product on a computer readable memory for controlling a computer apparatus to process electronic, said computer program product comprising:

mail fetching logic for fetching an electronic mail message for a user from a first mail server, said computer apparatus interacting as a first mail client with said first mail server;

mail filtering logic for identifying at least one predetermined characteristic within said electronic mail message that is indicative of said mail message being unwanted by said user so as to identify said electronic mail message as either a wanted electronic mail message or an unwanted electronic mail message;

mail storage logic for controlling storage of at least wanted electronic mail messages identified by said mail filtering logic; and

mail delivery logic responsive to a mail delivery request from a second mail client for delivering wanted mail for said user from said stored mail to said second mail client, said computer apparatus interacting as a second mail server with said second mail client.

23. Apparatus for processing electronic mail substantially as hereinbefore described with reference to the accompanying drawings.

24. A mehtod of processing electronic mail substantially as hereinbefore described with reference to the accompanying drawings.

The Patent Office

INVESTOR IN PEOPLE

Application No:	GB 9903672.5	**Examiner:**	B.J.SPEAR
Claims searched:	1-24	**Date of search:**	14 September 1999

Patents Act 1977
Search Report under Section 17

Databases searched:

UK Patent Office collections, including GB, EP, WO & US patent specifications, in:

UK Cl (Ed.Q): H4P (PPA,PX)

Int Cl (Ed.6): G06F 17/60, H04L 12/58

Other: Online:WPI, EPODOC, JAPIO

Documents considered to be relevant:

Category	Identity of document and relevant passage		Relevant to claims
X	GB2328110A	(Mitel) Whole document, eg Figs. 1, 2,	1,9,16,18, 19,22 at least
XY	EP0886228A2	(Digital Equipment) Whole document, eg abstract, Figs. 2 ,9 and col. 14 ll 20-56.	1,2,5,7,9-12,16-22 at least
X	EP0838774A2	(Tumbleweed) Whole document, eg abstract, Fig. 15 and claim 2.	1,16,17,19 and 22 at least
Y	EP0813162A2	(Sun) Whole document, eg pp 8-10	2,3 at least
X	WO9837680A2	(Intervoice) Whole document, eg abstract and Figs. 1,3	1,2,5,9,16 ,18,19 and 22 at least

X	Document indicating lack of novelty or inventive step	A	Document indicating technological background and/or state of the art.
Y	Document indicating lack of inventive step if combined with one or more other documents of same category.	P	Document published on or after the declared priority date but before the filing date of this invention.
&	Member of the same patent family	E	Patent document published on or after, but with priority date earlier than, the filing date of this application.